MONA LISA: A Framework for Reasoning About Gel Electrophoresis of Nucleic Acids

Submitted to the Hawaii International Conference on System Sciences
Motifs in Biology: Analysis of Ambiguous Data
Biotechnology Computing Track

Michael Cook†

Michiel Noordewier†‡

Joshua Lederberg†

†*The Rockefeller University*

‡*Rutgers University*

This work was supported by DARPA grant #MDA972-91-J-1008
ARPA order #8145/02

Introduction

We are developing MONA LISA, a knowledge-based system in the domain of gel electrophoresis. The purpose of this paper is to introduce this problem to computer scientists, and discuss our work to date on representations and algorithms for some aspects of this area.

A branch of modern experimental molecular biology attempts to determine the behaviour of organisms by observing the alteration of the information-containing molecules DNA and RNA. A principle tool for the characterization of these nucleic acids is their migration pattern in an gel under the influence of an electric field. Many experimental results, however, lend themselves to multiple interpretation. We therefore have found it useful to formally enumerate the possible results, each of which corresponds to a hypothesis about the events which have transpired in the experiment.

The goal of the MONA LISA system is to generate and evaluate hypotheses about in molecular biology experiments. This is accomplished in two steps:

1. DNA and RNA sequences which are the plausible products of the experiment are systematically generated,
2. The predicted gel electrophoretic behaviour of the putative molecules is compared with the actual gel results to eliminate hypotheses.

Input to the program is a set of descriptions defining:

- nucleic acid molecules
- protein factors
- experimental conditions
- gel electrophoresis parameters and results

Output from the program is set of hypotheses, where a hypothesis is defined to be an assignment of DNA and RNA molecules to bands which are observed on the gel.

The existence of a computational system to examine gel-electrophoretic experiments is expected to be of utility to a biologist because it addresses a problem which arises continuously in the laboratory. Although the generation of a specific assignment of nucleic acid molecules to bands is not difficult, the number of potential such assignments grows combinatorially, and is therefore difficult to exhaustively enumerate manually.

The interpretation of gel electrophoresis data is a good challenge for knowledge-based programming. First, its achievement would be useful to researchers — it is not a toy problem. Second, its solution requires further research into important areas such as knowledge representation and qualitative reasoning.

Gel Electrophoresis

Gel electrophoresis is one of the most widely used techniques today in molecular biology. It is based on the fact that most biological macromolecules are electrically charged and will therefore move in an electric field. "This property can be used to determine molecular weights, to distinguish molecules by virtue of their net charge or shape,...,and to separate different molecular species quantitatively" [5].

The many applications of gel electrophoresis include: DNA sequencing, Southern transfers, restriction mapping, separating DNA by molecular weight, or by shape due to conformation (e.g. supercoiling), detection of recombinant plasmids in a cloning experiment, analysis of unknown mixtures, peptide analysis, etc. Many different types of gels are in use: capillary gel electrophoresis, pulsed field, temperature gradient, 2D agarose, 2D polyacrylamide, transverse gradient, and denaturing gradient gel electrophoresis, to name a few (for an overview of electrophoresis see [4, 2, 1]).

The detailed theory of electrophoresis is "highly complicated and at present incomplete" [5]. Therefore, in practice, researchers use empirically determined heuristics to establish the conditions used in gel experiments, as well as in their interpretation. The problem of interpreting a gel is not a "well-formed" problem. The exact goal varies from one context to another, depending on the level of resolution of the data and the goals of the experiment. Also, the problem states are not discrete, and the operators used to move

between states are not obvious. At this point, gel data is often not that well quantified. For instance, the total amount of material involved in the experiment may not be known, and so the obvious constraint imposed by the conservation of mass is not available. Extra material which cannot be accounted for is often simply ignored.

However, the technique of gel electrophoresis is remarkably useful to biologists, and there is a body of knowledge that qualifies as expertise. For these reasons we decided that gel electrophoresis experiments are a good domain in which to build an expert system for use by researchers to assist in the design and interpretation of gel experiments.

Because of the interests in our laboratory, we have focused on one-dimensional (separation) gels with nucleic acids. The main other category of gels is 2-D gels, and some work has been done in the automatic scanning, matching, and interpretation of such gels [7, 9].

We are more interested in automatic reasoning and knowledge representation than image processing and data analysis, although in a complete gel system all these functions would be integrated.

Example

In our lab a significant amount of time is spent reasoning about gels. Often a gel is run in order to confirm that an experiment has produced the expected result. For example, if four different species of DNA are expected as the result of a certain procedure, one would expect four bands to appear on a gel. If only three bands appear, the question "what happened to the fourth band?" naturally arises. Several explanations are possible, and they are generated by members of the lab, and discussed for plausibility. Each such explanation is a candidate hypothesis which often suggests follow up experiments to test it. This is a situation in which the systematic enumeration of possibilities based on a knowledge base of facts and heuristic rules could be useful to the researcher.

Expert Systems

Because of the focus on generating hypotheses to fit data, we have been very much influenced by the DENDRAL paradigm, Plan-Generate-Test [3, 8]. DENDRAL was an early expert system designed to interpret mass spectroscopy experiments.

DENDRAL's task of inferring the structure of a molecule from its mass spectrum is analogically replaced by that of inferring the set of molecular species loaded into a gel from the pattern on the gel. However, a gel is run in many different contexts and this distinguishes it from the situation DENDRAL handles, which covers a standardized instrumental paradigm. Thus, the knowledge base for a gel system is richer and more diverse.

Outline of Paper

We present a data structure for representing gels, in order to concretize the sub-class of gels we are considering. Our basic model of experiments involving gels is: an experiment E is performed on an analyte N, resulting in a set of molecular species S; these species are run on a gel G; which is then interpreted as a set of bands B. Diagrammatically,

$$E : N \xrightarrow{1} S \xrightarrow{2} G \xrightarrow{3} B$$

This structures our discussion, and suggests a general framework for reasoning about nucleic acid gels. Each arrow in the above sequence suggests a different point of view on the problem.

1. The passage from analyte to a set of molecular species is modelled by rules of the form:

   $$\textit{Reagent: Nucleic Acid} \longrightarrow \textit{Products}$$

   The reagent could be an enzyme, or possibly null. We are building an "enzymatic production system" consisting of such rules, which is discussed in the section on a language for nucleic acids operations.

2. The passage from a set of molecular species to a migration pattern on a gel is a step involving the theory of gel electrophoresis, which is little understood, and not directly addressed in this paper. The expertise in this domain can be modeled by rules of the form:

$$\textit{Nucleic Acid} \times \textit{Gel} \longrightarrow \textit{Migration Distance}$$

Different types of nucleic acid and gel parameters result in different migration behaviours, some of which are at best empirically known, many of which are not.

Some very basic heuristics from this domain have informed our hypothesis generation algorithm, and as we pull more rules through the knowledge acquisition bottleneck, they will be used in a way discussed below. The most basic rule of thumb is one we have named "Monotonicity:"

Rule of Monotonicity: If nucleic acid A is longer than nucleic acid B, it will migrate more slowly.

This rule has many exceptions and ramifications which form much of the lore in this domain.

3. Usually the kinds of gels we are studying are described as a series of "bands" - discrete areas of co-migrating material which often but not always consist of homogeneous molecules. The passage from the gel to this more abstract description is accomplished by eye as an act of perception. Our current approach is to take the bands as a given and reason about them, but we believe that a gel can be scanned, and in most cases, bands isolated which correspond to what is perceived (in difficult cases, a band can be resolved by running a gel under different conditions).

Thus, in this paper, steps 2 and 3 are collapsed into one step,

$$S \longrightarrow B$$

In this context, we present a generator of hypotheses, where a hypothesis is defined to be an assignment of species to bands, that is a function

$S \longrightarrow B$. In addition, we present a scoring function which ranks hypotheses in order of likelihood.

What is a Gel?

This section describes a structure for representing a gel experiment. A *gel experiment* is: $\{G, P\}$

1. G = *Global gel parameters*
   - concentration of matrix (polyacrylamide, agarose, etc.)
   - physical dimensions of gel
   - applied voltage
   - length of run
   - goal (purify, analyze, separate, etc.)
2. P = *A set of lanes*, each with the following structure:

   lane = (experimental conditions, data)

   where *experimental conditions* is a vector of the ingredients that have been loaded into the *lane*, and *data* is a set of values representing the amount of material at distances $d_1, d_2, d_3, ..., d_k$ from the well.

A hypothetical example is shown in Figure 1: in the diagram each column is a *lane*, and the global parameters are at top. Also, in each data value we simply indicate the presence or absence of a band, rather than any quantitative amount.

The main point of this example is as follows: often lanes of one experiment are compared. They form what Simon calls a "data-cluster". We want to relate differences in the experimental conditions to differences in the data, in a manner similar to Simon's BACON [6]. Many gel interpretations are based on a comparison of lanes in one gel, since differences in the gel material from gel to gel makes it difficult to compare one gel to another. Very often there is a marker lane, which contains a material whose migration characteristics are well known, and this lane is used to calibrate the parameters of the gel in order to interpret the other lanes.

GEL #1 - effects of UV radiation

PAGE 7%
15 cm x 20 cm
V = 1000
t = 2 hrs.
goal = compare

Lane	1	2	3	4	5	6
Experimental Conditions:						
Mg++,	+	+	+	-	-	-
RNAp,	[x]	[x]	[x]	[x]	[x]	[x]
pBS DNA,	[y]	[y]	[y]	[y]	[y]	[y]
UV Light,	+	+	+	+	+	+
Time,	5	10	20	30	40	60
DATA:						
	-	-	-	-	-	-
	-	-		-	-	
	-		-	-		-
		-	-		-	
	-	-	-	-	-	-
	-	-		-	-	-
	-	-	-	-		-
		-	-	-	-	
	-	-	-	-	-	-
	-				-	-
		-	-			
	-	-	-	-	-	-
	-		-	-	-	-

Figure 1: Hypothetical Gel Experiment

A Little Language for Nucleic Acids Research

As previously described, the passage from analyte to a set of molecular species is modelled by rules of the form:

Reagent: Nucleic Acid $\longrightarrow$ *Products*

The reagent can be an enzyme, or possibly null. Our "enzymatic production system" consists of such rules, which models the common transformations applied to nucleic acids by biological and physical reagents. For example, we have included rules which describe the results of application of a restriction enzyme or a DNA polymerase.

The antecedents of these rules match data structures which describe individual species of DNA and RNA molecules. To facilitate the use of such rules, we describe an enhanced string language for representing nucleic acids. The language includes conventions for representing single stranded molecules, double stranded molecules, RNA, DNA, and RNA/DNA hybrids; for distinguishing between the two strands of a double stranded molecule, and for keeping track of the 5' to 3' orientation of a sequence. Our formalism supports operations representing the action of basic enzymes used in genetic engineering. We have implemented a parser in PROLOG for this syntax.

The "full" representation is a unambiguous representation of a double stranded DNA molecule, for instance:

```
5' - gaattcaaa - 3'
3' - cttaag... - 5'
```

The dots are place holders, indicating the absence of nucleotides. It is implied that both strands are covalently bonded, and hydrogen bonded with each other. Our goal is to write down rules to describe nucleic acid experiments in a one-dimensional way which is easily understandable, and reflects the informal way we describe these situations at lab meetings, but is formal enough to allow automatic reasoning and the establishing of provable properties.

First we give the conventions for representing molecules, and then we give some rules describing enzymatic reactions on nucleic acid molecules.

Convention 1: The left to right direction always represents 5' to 3'.

Convention 2: Lower-case characters refer to ssDNA.

Convention 3: Upper-case characters refer to dsDNA

Convention 4: Characters in quotes are literal nucleotide specifications:

```
'AGC', 'gaag'
```

Convention 5: Characters outside quotes are variables specifying strings of nucleotides.

Convention 6: The complement operator is " ~ ", and refers to the biological complement of a sequence, i.e. the sequence of the complementary strand (if the molecule is double stranded). It can only refer to single stranded sequence.

Examples illustrating the first five conventions:

```
'AGC'   <==>  5'-agc-3'
              3'-tcg-5'

'agc'   <==>  5'-agc-3'

~'agc'  <==>  5'-tcg-3'
```

Convention 6: A caret " ^ " or underscore " _ " following any expression indicates that the lower-case characters in the expression are on the upper or lower strand, respectively, (e.g. b^ , 'gaa'_ , 'AAggg'^).

Some examples:

```
'AAggg'^ <==> 5'-aaggg-3'
              3'-tt   -5'

'AAggg'_ <==> 5'-aaccc-3'
              3'-tt   -5'
```

```
'gaa'_   <==> 5'-ctt-3'

'gaa'_   <==> ~'gaa'^
```

Convention 7: Nucleotides within a string are indexed by optional parentheses following the string variable:

```
'AAGCTTG'(4,7)  <==>  5'-cttg-3'
                      3'-gaac-5'
```

Convention 8: (Convention 1 revisited)

Sequences are written in a "canonical" 5' to 3' direction. Single stranded regions are written as the sequence they would be if paired on the "upper" strand:

```
'AAggg'^   <==>   5'-aaggg-3'
                  3'-tt   -5'

'aattC'_   <==>   5'-    c-3'
                  3'-ttaag-5'
```

Convention 9: A DNA molecule is specified as one or more segments separated by commas within square brackets:

```
[ R, 'GAATTC', S ]
```

Convention 10: An RNA molecule is specified as one or more segments separated by commas within curly brackets:

```
{ R, 'GAATTC', S }
```

Convention 11: A DNA/RNA hybrid molecule is specified as a post-fix notation on a segment within square or curly brackets indicating the composition of one of the strands (modifying the nucleic acid type specified:

```
{ 'GAATTC':D }    <==>   5'-gaattc-3'   DNA
                         5'-cttaag-3'   RNA

{ 'GAATTC':d }    <==>   5'-gaattc-3'   RNA
                         5'-cttaag-3'   DNA

[ 'GAATTC':R ]    <==>   5'-gaattc-3'   RNA
                         5'-cttaag-3'   DNA

[ 'GAATTC':r ]    <==>   5'-gaattc-3'   DNA
                         5'-cttaag-3'   RNA
```

A molecule which mixes DNA and RNA on the same backbone can be specified as above for hybrid molecules:

```
                              DNA    RNA
                              ------
[ X, Y:R ]        <==>   5'-xxxxxxyyyyyy-3'
                         3'-xxxxxxyyyyyy-5'
                            ------------
                               DNA
```

These eleven conventions allow the representation of a wide variety of nucleic acid molecules.

Rules for Enzymatic Manipulation of Nucleic Acid Molecules

Utilizing the above conventions, we can describe the actions of enzymatic agents on DNA and RNA. The general approach is to match the antecedent of a rule to a description of a set of nucleic acid molecules, binding the sequence and structural properties to variables in the left hand side of the rule. The rule then acts as a production, to create the description of a product set of molecules.

Rules have the form:

antecedent molecules $\longrightarrow$ consequent molecules

Examples of rules are found in Figure 2. Hypotheses about the behaviour of processes on informational molecules *in vitro* are thus confirmed or denied by examining the creation or modification of nucleic acids.

We note that there are a number of relevant biological processes which are not addressed yet, but should be handled by straightforward extension of the rule syntax we have described. These include nicking reactions and circular molecules.

Other extensions will include more unusual conformational states of DNA and RNA, such as supercoiling, or the formation of triplex molecules and non-canonical hybrids, which are coming under increasing scrutiny from the molecular biology community.

EcoR1 endonuclease:

$[\ R, \text{'GAATTC'}, S\ ] \longrightarrow [\ R, \text{'Gaatt'}\_\ ] + [\ \text{'aattC'}\hat{}\ , S\ ]$

DNA polymerase (progressive):

$[\ A, b\_(1,n)\ ] \longrightarrow [\ A, B(1), b\_(n,2)\ ]$

DNA polymerase (complete):

$[\ A, b\_\ ] \longrightarrow [\ A, B\ ]$

DNA Ligase (sticky ended molecules):

$[\ X, s\_\ ] + [\ s\hat{}\ , Y\ ] \longrightarrow [\ X, S, Y\ ]$

RNA Polymerase (intermediate state):

$[\ X, P, Y\ ] \longrightarrow [\ x\hat{}\ , p\hat{}\ , y\hat{}\ ] + [\ x\_, p\_, Y{:}R\ ]$

RNA Polymerase (final state):

$[\ X, P, Y\ ] \longrightarrow [\ X, P, Y\ ] + \{\ R\ \}$

DNA Ligase (blunt ended molecules):

DNA ligase: $[X] + [Y] \longrightarrow [X,Y]$

Reverse Transcriptase:

$\{x\} \longrightarrow \{X{:}d\} \longrightarrow [\ \tilde{x}] \longrightarrow [X]$

Exonuclease:

$[X(1,n)] \longrightarrow [X(1,n\text{-}1)]$

Annealing:

$[x] + [\ \tilde{x}] \longrightarrow [X]$

Figure 2: Rules for Enzymatic Manipulation of Nucleic Acids

An Hypothesis Generator for the Assignment of Molecular Species to Bands

We consider experiments with the following structure:

1. Run a reaction involving nucleic acids, resulting in a set of molecular species $S = s_1, s_2, ..., s_k$.
2. Load the resulting material in a well in a gel, and electrophorese.
3. Visualize the material in the gel, by autoradiography, staining, or some other procedure.

Schematically, we have

$$E : N \longrightarrow S \longrightarrow G \longrightarrow B$$

where E is the experiment, N the analyte, S is the set of molecular species resulting, G is the gel, and B is a set of bands that are perceived. Given the experimental results G and the input data to the experiment E we want computer assisted hypothesis formation about the content of the gel.

We attack a very simple case first, one in which we "know" S and B. That is, we assume a well-defined set of distinct molecules, of differing molecular weights, S. We also assume a well-defined set of distinct "bands" on a gel, $B = b_1, b_2, ..., b_n$. In this context, a "hypothesis" is an assignment of species to bands, that is, a function $f : S \rightarrow B$.

In this simplified setting, we can reason about the set of all hypotheses, and generate systematically a reasonably constrained subset of them. The first observation is that if all mappings $f : S \rightarrow B$. are considered, the set of hypotheses is k^n. As an illustration of how reasonable constraints can dramatically prune the search space, notice that if $k = n$, and we focus attention on one to one functions, the size of the resulting set is $n!$. (This corresponds to assuming that each species of molecule can only appear in one band, and to assuming that two species of molecules do not co-migrate. These assumptions do not always hold, but are not unreasonable.)

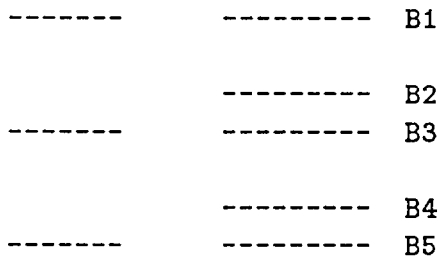
To further cut down the size of the search space, we impose the further constraint of monotonicity, that is, we assume that S is sorted by size, that B is sorted by migration distance from the well at the top of the lane, and that mappings from S to B are monotonic. In this case, with $k = n$, there is exactly one hypothesis that fits the data, the unique 1 to 1 function $f : S \rightarrow B$ that maps decreasing weights into faster migrating bands.

We consider next the situation in which the number of bands and the number of molecular species differ. There are two cases to consider:

1. $|S| < |B|$ — less species than bands;
2. $|S| > |B|$ — more species than bands;

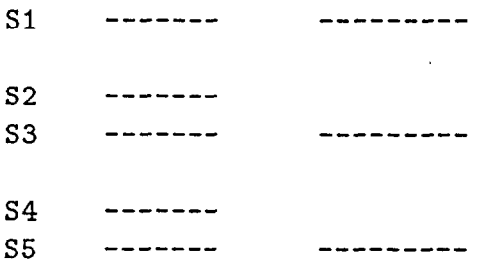
In each case, we would like a generator of hypotheses, where each hypothesis satisfies the monotonicity requirement. (A situation which did not satisfy this condition is a good candidate for what is loosely termed "anomalous migration.") Before we analyze the general case, an example of case 1) and case 2) should clarify the discussion.

Example: 5 bands, 3 species:



There are $\binom{5}{3}$ mappings. Note that if the species are assumed to occupy bands 1, 3, and 5, then B2 can be hypothesized to be material from either B1 or B3; and B4 can be hypothesized to be material from either B3 or B5.

EXAMPLE: 3 bands, 5 species:



There are $\begin{pmatrix} 5 \\ 3 \end{pmatrix}$ mappings. Note that if the bands are assumed to be occupied by species 1, 3, and 5, then S2 can be hypothesized to co-migrate with either S1 or S3; and S4 can be hypothesized to co-migrate with either S3 or S5.

In case 1), there are $\begin{pmatrix} n \\ k \end{pmatrix} \sim n^k$ one to one functions from S to B, the number of ways of choosing which k bands are hit by elements of S. After the target bands are chosen, one must still account for the remaining bands. For each such "remainder" there are two possibilities, within the constraints of monotonicity: either it is material from the band above it, or it is material from the band below it. If there is no band above it, then we assume it is from the band below, and if there is no band below it we assume it is from the band above it. If the remaining bands are all interior (not the top band or the bottom band), the number of hypotheses is:

$$\begin{pmatrix} n \\ k \end{pmatrix} * (n-k) * 2$$

This formula can be easily adjusted for a case in which a remainder band is at an extreme position on the gel.

In case 2), there are $\begin{pmatrix} k \\ n \end{pmatrix}$ functions which map onto B, this being the number of ways of choosing n species that have been collapsed into neighboring bands. A further complication is the question as to which neighboring bands they have collapsed to. This is a question of which bands have co-migrated with which (again, within the constraints of monotonicity). This situation is entirely analogous to the above, and the formula is the same.

To generate all the hypotheses associating molecular species with bands within the above framework, we can first generate a mapping, and then for each mapping, generate the $2 * (n - k)$ assignment of missing bands, or missing species. So the first question at hand is: find an algorithm to systematically generate all subsets of k elements in an n element set. We present an algorithm in the next section.

Algorithm for the Generation of All Subsets of Size k in a Set of Size n

Recall the recursion relation:

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

We use this to generate all n bit numbers with exactly k bits equal to 1. Once we have done this, it is clear how to associate this with subsets of an n element set.

Let $T = \{n$ bit numbers with exactly k bits turned on$\}$

The observation used is simply that this set consists of two subsets, odd numbers whose last bit is 1; and even numbers, whose last bit is 0. The first set has $k - 1$ of its first $n - 1$ bits turned on, the second set has k of its first n bits turned on. Thus if we define O and E by:

- $O = \{n - 1$ bit numbers with exactly $k - 1$ bits turned on $\}$
- $E = \{n - 1$ bit numbers with exactly k bits turned on$\}$

Then if $T = \{n$ bit numbers with exactly k bits turned on$\}$

$$T = \{2 * O + 1\} \cup \{2 * E\}$$

Now we must specify the base of the recursion, which we do as follows:

For $\binom{n}{n}$, we return the number whose binary representation is n 1's. For $\binom{n}{0}$, we return the n bit number all of whose bits are off, i.e., zero.

This algorithm has been coded in LISP, and can be used to generate all constrained hypotheses in the above described context of gel experiments.

It is reasonable to discuss at this point the various heuristics that can be used to rank these hypotheses. For instance, middle bands are often given more weight than bands at either end of the gel. Also, more intense bands are given often given more weight. However, it is interesting that important discoveries have been made by focussing on faint bands - for example, ribozymes, and the reverse transcriptase activity of Taq polymerase.

Ranking of Hypotheses

One of the very interesting aspects of this project is a chance to study multiple levels of interacting hypotheses. A typical gel discussion might have the following "hypothesis structure":

At a top level, there is a hypothesis about the migration of nucleic acids, for example:

- Hypothesis 1: If I plot the migration of nucleic acids of known molecular weights on semi-log paper (weights against distance), the curve can be fitted with a cubic polynomial. Given migration distances for the unknown material, I can then use this curve to estimate its molecular weight.

  Remark: This hypothesis is open to question, because there is always the possibility of anomalous migration, due to some condition that has not yet been documented.

  Given this working hypothesis, working hypotheses about the existence of species loaded into the wells are formed:

- Hypothesis 2: Species s_1, s_2, and s_3 have resulted from the experiment performed and are present in the gel.

- Hypothesis 3: The above species have molecular weights of w_1, w_2, and w_3 respectively.

Remark: The second two hypotheses are also often rethought during the course of a discussion.

Finally, in the context of the above hypotheses, hypotheses about the association of species with bands may be formed, which is the level of discussion addressed in the previous section. However, the existence of these multiple levels of hypotheses, the way they interact in practice, and the way they are modified and adjusted in the course of a typical discussion among experts, is, ultimately, the complex knowledge structure we hope to formalize.

In this section we discuss only the last mentioned level of hypothesis formation, and present one possible measure of the "likelihood" of such an hypothesis.

For an hypothesis which takes the form of a list of pairs:

$$(Weight_i, Distance_i), i = 1, ..., k$$

with descending weights and ascending distances, we define a vector consisting of ratios of successive differences as follows:

$$R_i = \frac{log(w_{i+1}) - log(w_i)}{d_{i+1} - d_i}, i = 1, ..., k-1$$

Then define the "variation" of a hypothesis as the maximum distance between these ratios:

$$\overset{max}{i,j} \mid R_i - R_j \mid$$

In the absence of anomalous migration, a mapping from weights to bands is therefore more likely if it has less variation, i.e. the best hypothesis is gotten by choosing the vector with the least variation.

An example should clarify this proposed rule for ranking hypotheses.

Example: Given DNA fragments of 10, 20 and 30 base pairs, and a gel with bands at 2cm, 4cm, 4.1cm, and 6cm from the origin, there are $\binom{4}{3} = 4$ hypotheses about which bands contain which species:

```
|
1)      (30 bp, 2 cm)
        (20 bp, 4 cm)
        (10 bp, 6 cm)

        R = (5,5)
        max = 0

2)      (30 bp, 2 cm)
        (20 bp, 4.1 cm)
        (10 bp, 6 cm)
```

```
        R = (10/2.1,10/1.9) = (4.76,5.26)
        max = .50

3)      (30 bp, 4 cm)
        (20 bp, 4.1 cm)
        (10 bp, 6 cm)

        R = (10/.1,10/1.9) = (100, 5.26)
        max = 94.74

4)      (30 bp, 2 cm)
        (20 bp, 4 cm)
        (10 bp, 4.1 cm)

        R = (10/2,10/.1) = (5, 100)
        max = 95
```

Thus, the ranking in this case is:

1. is the most likely
2. is the next most likely
3. is the next most likely
4. is the least likely.

Of course, in cases for which there is no good guess as to the sizes of the fragments, this rule is not applicable.

This area is complex, and is a focus of our current research. We anticipate finding different methods for ranking hypotheses, depending on the granularity of the data, the confidence factors associated with various data, and the particular goal of the experiment at hand.

The above discussion was based on knowing S and B. This is often not a fully realistic assumption in the world of gel electrophoresis. In reality, the process of going from an experiment to a set of molecular species is fraught with unknowns; and this aspect of modeling is addressed in our "enzymatic production system".

Summary

The process of thinking about gels as we have observed it, exhibits the following pattern:

1. Look at the gel, G, and discern its significant features: its bands, B, their intensity, thickness, and number, areas of smear, and any anomalies.

2. Consider the experiment, and hypothesize a set of species that are expected to appear.

3. Generate hypotheses about the association between molecular species and bands - and rank them according to "expert" knowledge.

4. Often, rethink the expected species, generating new hypotheses in the light of discussion; and rethink the description of the bands in the gel.

5. Finally, most gel discussions end with a suggestion for what experiment or experiments would be valuable to perform next, in order to resolve remaining ambiguities.

Once a set of species, S and a set of bands, B, are postulated, hypotheses about their possible associations $h : S \rightarrow B$ are enumerated with a simple generator, and ranked according to user imposed heuristics and criteria.

References

[1] *Gel electrophoresis of nucleic acids: a practical approach.* Oxford: IRL Press, 1982.

[2] Anthony T Andrews. *Electrophoresis: theory, techniques, and biochemical and clinical applications.* Oxford: Clarendon Press, 1981.

[3] Bruce G. Buchanan, G. L. Sutherland, and Edward A. Feigenbaum. *Heuristic DENDRAL: A Program for Generating Explanatory Hypotheses in Organic Chemistry*, volume 4, pages 209–254. Edinburgh University Press, 1969.

[4] Andreas Chrambach. *The practice of quantitative gel electrophoresis.* Deerfield Beach FL: VCH Publishers, 1985.

[5] David Michael Freifelder. *Physical biochemistry: applications to biochemistry and molecular biology.* San Francisco: W. H. Freeman, 1976.

[6] Pat Langley, Herbert A. Simon, Gary L. Bradshaw, and Jan M. Zytkow. *Scientific Discovery: Computational Explorations of the Creative Processes.* MIT Press, Cambridge, Massachussetts.

[7] P.F. Lemkin and L.E. Lipkin. GELLAB: A computer system for 2d gel electrophoresis analysis I,II. *Computers and Biomedical Research*, 14:355–380.

[8] Robert Lindsay, Bruce G. Buchanan, Edward A. Feigenbaum, and Joshua Lederberg. *Applications of artificial intelligence for organic chemistry: the DENDRAL Project.* New York: McGraw-Hill, 1980.

[9] Mark J. Miller. Computer analysis of two-dimensional gels: Semi-automatic matching. *Clinical Chemistry*, 28(4):867–875, 1982.